Simple linear regression (continued) Stat 564
9-28-17

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad E[\varepsilon_i] = 0 \ \forall i$$

1

$$V[\varepsilon_i] = \sigma^2 \ \forall i$$

$\varepsilon_i, \varepsilon_j$ are independent $\forall i \neq j$

Find the properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

2

$$E[\hat{\beta}_1] = E\left[\frac{S_{xy}}{S_{xx}}\right]$$

$$= \frac{1}{S_{xx}} E[S_{xy}]$$

$$= \frac{1}{S_{xx}} E\left[\sum_{i=1}^{n}(x_i - \bar{x})y_i\right]$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x}) E[y_i]$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x}) E[\beta_0 + \beta_1 x_i + \varepsilon_i]$$

③

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i)$$

$$= \frac{1}{S_{xx}} \Big[\beta_0 \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{0} + \beta_1 \underbrace{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}_{S_{xx}} \Big]$$

$$= \frac{1}{S_{xx}} [\beta_1 S_{xx}] = \beta_1$$

$\therefore \hat{\beta}_1$ is an unbiased estimator of β_1

④

$$Var(\hat{\beta}_1) = V\left(\frac{S_{xy}}{S_{xx}}\right)$$

$$= \frac{1}{S_{xx}^2} V(S_{xy})$$

$$= \frac{1}{S_{xx}^2} V\left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i\right)$$

$$= \frac{1}{S_{xx}^2} \Big[\sum_{i=1}^{n} (x_i - \bar{x})^2 V(y_i) + \sum_{i<j}\sum (x_i - \bar{x})(x_j - \bar{x}) \cancel{Cov(y_i, y_j)}^{\,0} \Big]$$

because $\varepsilon_i, \varepsilon_j$ are indep

⑤

$$= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 V(\beta_0+\beta_1 x_i+\varepsilon_i)$$

$$= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \underbrace{V(\varepsilon_i)}_{\sigma^2}$$

$$= \frac{\sigma^2}{S_{xx}^2} S_{xx} = \frac{\sigma^2}{S_{xx}}$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{(n-1)s_x^2}$$

Since $\hat{\beta}_1$ was unbiased & $\lim_{n\to\infty} V(\hat{\beta}_1) = 0$, $\hat{\beta}_1$ is <u>consistent</u>

⑥

What about $\hat{\beta}_0$?

$$E[\hat{\beta}_0] = E[\bar{y} - \hat{\beta}_1\bar{x}] = E[\bar{y}] - \bar{x}E[\hat{\beta}_1]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] - \bar{x}\beta_1$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(\beta_0+\beta_1 x_i+\varepsilon_i) - \bar{x}\beta_1$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i + 0) - \bar{x}\beta_1$$

$$= \frac{1}{n}\left(n\beta_0 + \beta_1 \Sigma x_i\right) - \bar{x}\beta_1 = \beta_0$$

$\therefore \hat{\beta}_0$ is also unbiased

⑦

$$V[\hat{\beta}_0] = V[\bar{y} - \hat{\beta}_1 \bar{x}]$$

$$= V\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})y_i\right]$$

$$= V\left[\sum_{i=1}^{n} \underbrace{\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right)}_{k_i} y_i\right]$$

$$= \sum_{i=1}^{n} k_i^2 \underbrace{V[y_i]}_{\sigma^2} + 0$$

$$= \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right)^2$$

⑧

$$= \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} + \frac{\bar{x}^2(x_i - \bar{x})^2}{S_{xx}^2} - \frac{2\bar{x}(x_i - \bar{x})}{n S_{xx}}\right)$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{S_{xx}} - \frac{2\bar{x}}{n S_{xx}}\underbrace{\cancel{\sum(x_i - \bar{x})}}_{0}\right]$$

$$V(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right]$$

$\therefore \hat{\beta}_0$ is also a consistent estimator

We need to find an estimator of σ^2. ⑨

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i\right)^2$$

$$= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right]^2$$

⑩

$$= \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 + \hat{\beta}_1^2 (x_i - \bar{x})^2 - 2\hat{\beta}_1 (y_i - \bar{y})(x_i - \bar{x})\right]$$

$$SSE = S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}$$

$$= S_{yy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} - 2\left(\frac{S_{xy}}{S_{xx}}\right) S_{xy}$$

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy} - \hat{\beta}_1^2 S_{xx}$$

$$E(SSE) = E\left[S_{yy} - \hat{\beta}_1^2 S_{xx}\right]$$

$$= \underbrace{E\left[S_{yy}\right]}_{1} - S_{xx} \underbrace{E\left[\hat{\beta}_1^2\right]}_{2}$$

2: $V[X] = E[X^2] - (E[X])^2 \Rightarrow E[X^2] = V[X] + (E[X])^2$

$$E\left[\hat{\beta}_1^2\right] = V\left[\hat{\beta}_1\right] + \left(E\left[\hat{\beta}_1\right]\right)^2$$

$$= \frac{\sigma^2}{S_{xx}} + \beta_1^2$$

1: $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$$= \sum_{i=1}^{n} \left(\cancel{\beta_0} + \beta_1 x_i + \varepsilon_i - \left(\cancel{\beta_0} + \beta_1 \bar{x} + \bar{\varepsilon}\right)\right)^2$$

$$= \sum_{i=1}^{n} \left[\beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\right]^2$$

$$= \sum_{i=1}^{n} \left[\beta_1^2 (x_i - \bar{x})^2 + (\varepsilon_i - \bar{\varepsilon})^2 + 2\beta_1 (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})\right]$$

13

$$S_{yy} = \beta_1^2 S_{xx} + \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 + 2\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})$$

$$E[S_{yy}] = \beta_1^2 S_{xx} + E\left[(n-1)s_\varepsilon^2\right] + 2\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})\overset{0}{\cancel{E(\varepsilon_i - \bar{\varepsilon})}}$$

$$= \beta_1^2 S_{xx} + (n-1)\sigma^2$$

$$E(SSE) = 1 - S_{xx}\, 2$$

$$= \beta_1^2 S_{xx} + (n-1)\sigma^2 - S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right)$$

$$= (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

14

So $E\left(\frac{SSE}{n-2}\right) = \sigma^2$

<u>Defn</u>: $MSE = \frac{SSE}{n-2} = \hat{\sigma}^2$

"mean squared error"

The MSE is an unbiased estimator of σ^2

Summary:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$$

$$E[\hat{\beta}_1] = \beta_1, \quad E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\sigma}^2] = \sigma^2$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \quad V[\hat{\beta}_0] = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

$$\hat{V}[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{S_{xx}}, \quad \hat{V}[\hat{\beta}_0] = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

$\hat{\sigma}/\sqrt{S_{xx}}$ and $\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$ are the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$

2.4 *PANKRATZ · Forecasting with of least-squares estimat aid="5N5AR">Table B.3 presents data on the gasoline mileage performance of 32 different automobiles.*

a. *Fit a simple linear regression model relating gasoline mileage y (miles per gallon) to engine displacement x_1 (cubic inches).*

b. *Construct the analysis-of-variance table and test for significance of regression.*

c. *What percent of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement?*

d. *Find a 95% CI on the mean gasoline mileage if the engine displacement is 275 in.^{3}*

e. *Suppose that we wish to predict the gasoline mileage obtained from a car with a 275-in.^{3} engine. Give a point estimate of mileage. Find a 95% prediction interval on the mileage.*

f. *Compare the two intervals obtained in parts d and e. Explain the difference between them. Which one is wider, and why?*

2.7 *The purity of oxygen produced by a fractional distillation process is thought to be related to the percentage of hydrocarbons in the main condensor of the processing unit. Twenty samples are shown below.*

a. *Fit a simple linear regression model to the data.*

b. *Test the hypothesis H_0: $\beta_1 = 0$.*

c. *Calculate R^2.*

d. *Find a 95% CI on the slope.*

e. *Find a 95% CI on the mean purity when the hydrocarbon percentage is 1.00.*

Purily (%)	Hydrocarbon (%)	Purily (%)	Hydrocarbon (%)
86.91	1.02	96.73	1.46
89.85	1.11	99.42	1.55
90.28	1.43	98.66	1.55
86.34	1.11	96.07	1.55
92.58	1.01	93.65	1.40
87.33	0.95	87.31	1.15
86.29	1.11	95.00	1.01
91.86	0.87	96.85	0.99
95.61	1.43	85.20	0.95
89.86	1.02	90.56	0.98